Reading for Mon (Sun night): 2.7

Today: a histogram

mean & median.

Histogram example. Directions say 5 buckets.

precision of data: rounded to 1 decimal place

min = 33.7 → start of 1st bucket (class)

max = 48.6

33.65

why? because 33.7 was rounded to 1 place so it really could fall between 33.65 & 33.75

$$\frac{\text{max} - \text{min}}{\#\text{ buckets}} = \frac{48.6 - 33.7}{5} = 2.98$$

↓ round up to nearest whole number

bucket width = 3

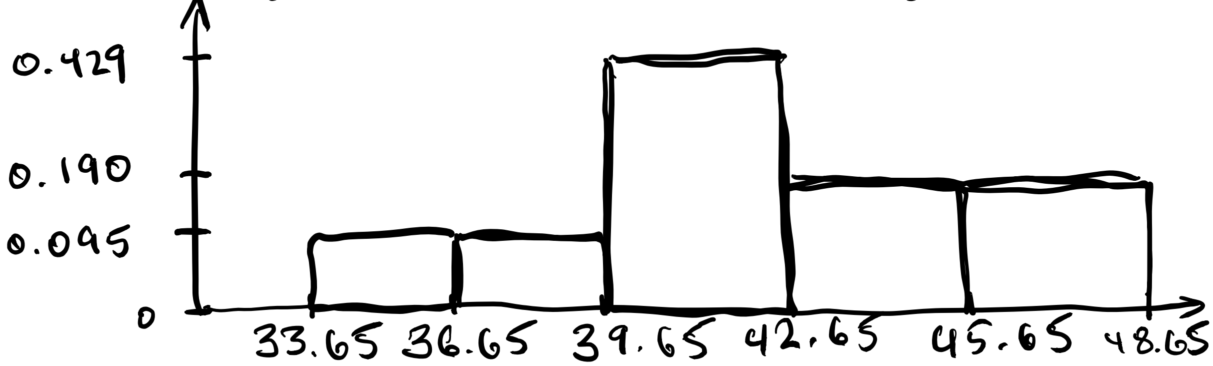
start at starting point

add bucket width

$rf = \frac{f}{n}$

Buckets	frequency f		relative frequency
33.65 – 36.65	2	÷21 =	0.095 = 9.5%
36.65 – 39.65	2	÷21 =	0.095
39.65 – 42.65	9	÷21 =	0.429
42.65 – 45.65	4	÷21 =	0.190
45.65 – 48.65	4	÷21 =	0.190
+	n = 21		

Histogram



Histogram example 2. Use 4 bars.

min = 73.231 $\longrightarrow$ start at ~~73.2~~.

$$\begin{array}{r} 73.231 \\ -\ 0.0005 \\ \hline 73.2305 \end{array}$$

why? because given data is rounded to 3 decimal places, the min 73.231 really

(1) falls anywhere between 73.2305 and 73.2315

(2) Our data has 3 decimal places, so if all the class endpoints have 4 decimal places, no data point will fall exactly on a endpoint.

max = 103.280

$$\text{class width} = \frac{\text{max} - \text{min}}{\#\text{ bars}} = \frac{103.280 - 73.231}{4} = 7.51$$

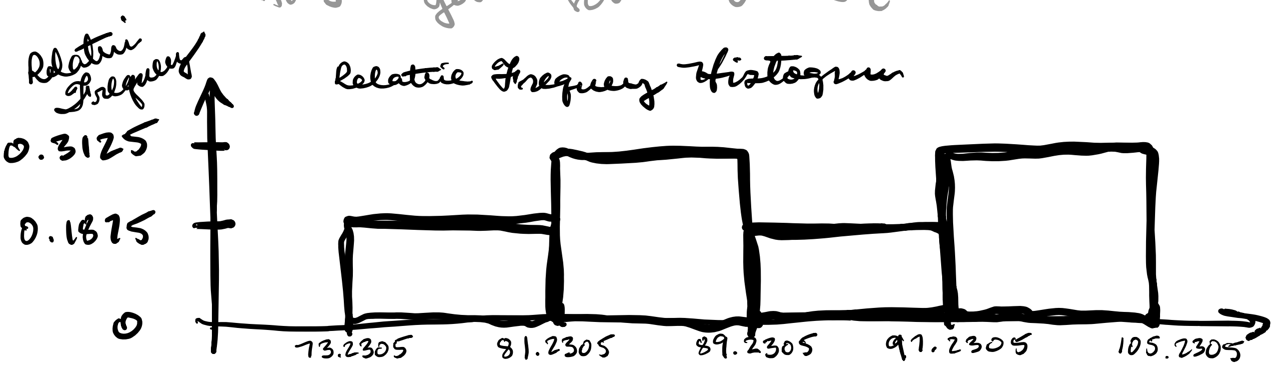
$\to$ round to 8

Classes	frequency		relative frequency
73.2305 – 81.2305	3	$\div 16 =$	0.1875
81.2305 – 89.2305	5		0.3125
89.2305 – 97.2305	3		0.1875
97.2305 – 105.2305	5		0.3125
	$n = 16$		

+8 +8 +8 +8

they will tell you what to round to or your choice

Relative Frequency Histogram



Try It Σ

2.24 The following data set shows the heights in inches for the boys in a class of 40 students.

66; 66; 67; 67; 68; 68; 68; 68; 68; 69; 69; 69; 70; 71; 72; 72; 72; 73; 73; 74
The following data set shows the heights in inches for the girls in a class of 40 students.
61; 61; 62; 62; 63; 63; 63; 65; 65; 65; 66; 66; 66; 67; 68; 68; 68; 69; 69; 69
Construct a box plot using a graphing calculator for each data set, and state which box plot has the wider spread for the middle 50% of the data.

Example 2.25

Graph a box-and-whisker plot for the data values shown.

10; 10; 10; 15; 35; 75; 90; 95; 100; 175; 420; 490; 515; 515; 790

The five numbers used to create a box-and-whisker plot are:

Min: 10
Q_1: 15
Med: 95
Q_3: 490
Max: 790

The following graph shows the box-and-whisker plot.

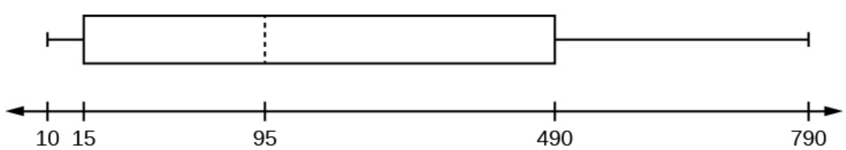


Figure 2.15

Try It Σ

2.25 Follow the steps you used to graph a box-and-whisker plot for the data values shown.

0; 5; 5; 15; 30; 30; 45; 50; 50; 60; 75; 110; 140; 240; 330

2.5 | Measures of the Center of the Data

The "center" of a data set is also a way of describing location. The two most widely used measures of the "center" of the data are the **mean** (average) and the **median**. To calculate the **mean weight** of 50 people, add the 50 weights together and divide by 50. To find the **median weight** of the 50 people, order the data and find the number that splits the data into two equal parts. The median is generally a better measure of the center when there are extreme values or outliers because it is not affected by the precise numerical values of the outliers. The mean is the most common measure of the center.

NOTE

The words “mean” and “average” are often used interchangeably. The substitution of one word for the other is common

Sample mean $\bar{x}$
population mean μ

practice. The technical term is "arithmetic mean" and "average" is technically a center location. However, in practice among non-statisticians, "average" is commonly accepted for "arithmetic mean."

When each value in the data set is not unique, the mean can be calculated by multiplying each distinct value by its frequency and then dividing the sum by the total number of data values. The letter used to represent the **sample mean** is an x with a bar over it (pronounced "x bar"): $\bar{x}$.

The Greek letter μ (pronounced "mew") represents the **population mean**. One of the requirements for the **sample mean** to be a good estimate of the **population mean** is for the sample taken to be truly random.

method 1

To see that both ways of calculating the mean are the same, consider the sample:
1; 1; 1; 2; 2; 3; 4; 4; 4; 4; 4

given sample

$$\frac{1+1+1+2+2+3+4+4+4+4+4}{11} = \frac{30}{11} = \boxed{2.7}$$

Sample mean $\bar{x} = 2.7$

method 2

$$\frac{1\times3+2\times2+3\times1+4\times5}{11} = \frac{30}{11} = \boxed{2.7}$$

(also known as weighted average)

method 3 → this is cool because you see that
"sum of (value) × (relative frequency)" = mean

$$1\times\frac{3}{11}+2\times\frac{2}{11}+3\times\frac{1}{11}+4\times\frac{5}{11} = \frac{30}{11} = \boxed{2.7}$$

values we see in data set

yes, method 3 is secretly the same as method 2

Example 2.26

AIDS data indicating the number of months a patient with AIDS lives after taking a new antibody drug are as follows (smallest to largest):
3; 4; 8; 8; 10; 11; 12; 13; 14; 15; 15; 16; 16; 17; 17; 18; 21; 22; 22; 24; 24; 25; 26; 26; 27; 27; 29; 29; 31; 32; 33; 33; 34; 34; 35; 37; 40; 44; 44; 47;
Calculate the mean and the median.

Solution 2.26

The calculation for the mean is:

$$\bar{x} = \frac{[3+4+(8)(2)+10+11+12+13+14+(15)(2)+(16)(2)+...+35+37+40+(44)(2)+47]}{40} = 23.6$$

To find the median, M, first use the formula for the location. The location is:

$$\frac{n+1}{2} = \frac{40+1}{2} = 20.5$$

Starting at the smallest value, the median is located between the 20th and 21st values (the two 24s):
3; 4; 8; 8; 10; 11; 12; 13; 14; 15; 15; 16; 16; 17; 17; 18; 21; 22; 22; 24; 24; 25; 26; 26; 27; 27; 29; 29; 31; 32; 33; 33; 34; 34; 35; 37; 40; 44; 44; 47;

$$M = \frac{24+24}{2} = 24$$

# of movies	Relative Frequency
0	$\frac{5}{30}$
1	$\frac{15}{30}$
2	$\frac{6}{30}$
3	$\frac{3}{30}$
4	$\frac{1}{30}$

Table 2.24

If you let the number of samples get very large (say, 300 million or more), the relative frequency table becomes a relative frequency distribution.

A **statistic** is a number calculated from a sample. Statistic examples include the mean, the median and the mode as well as others. The sample mean $\bar{x}$ is an example of a statistic which estimates the population mean μ.

Calculating the Mean of Grouped Frequency Tables

When only grouped data is available, you do not know the individual data values (we only know intervals and interval frequencies); therefore, you cannot compute an exact mean for the data set. What we must do is estimate the actual mean by calculating the mean of a frequency table. A frequency table is a data representation in which grouped data is displayed along with the corresponding frequencies. To calculate the mean from a grouped frequency table we can apply the basic definition of mean: *mean* = $\frac{data\ sum}{number\ of\ data\ values}$ We simply need to modify the definition to fit within the restrictions of a frequency table.

Since we do not know the individual data values we can instead find the midpoint of each interval. The midpoint is $\frac{lower\ boundary + upper\ boundary}{2}$. We can now modify the mean definition to be $Mean\ of\ Frequency\ Table = \frac{\sum fm}{\sum f}$ where f = the frequency of the interval and m = the midpoint of the interval.

This process I call the binned mean
bucketed mean

Example 2.30

same as in histogram exercise ~~(tabular ...)~~

A frequency table displaying professor Blount's last statistic test is shown. Find the best estimate of the class mean.

This example is for the whole class, so is a census, we look for population mean.

Frequency

Grade Interval	Number of Students
50–56.5	1
56.5–62.5	0
62.5–68.5	4
68.5–74.5	4
74.5–80.5	2
80.5–86.5	3
86.5–92.5	4
92.5–98.5	1

class width +6

Table 2.25

bucket mid point = $\frac{\text{top} + \text{bot}}{2}$

$\frac{(56.5+50)}{2} = 53.25$
$(56.5+62.5)/2 = 59.5$
$(62.5+68.5)/2 = 65.5$
$(68.5+74.5)/2 = 71.5$
$(74.5+80.5)/2 = 77.5$
$(80.5+86.5)/2 = 83.5$
$(86.5+92.5)/2 = 89.5$
$(92.5+98.5)/2 = 95.5$

+6 +6 +6

Solution 2.30

- Find the midpoints for all intervals

Grade Interval	Midpoint
50–56.5	53.25
56.5–62.5	59.5
62.5–68.5	65.5
68.5–74.5	71.5
74.5–80.5	77.5
80.5–86.5	83.5
86.5–92.5	89.5
92.5–98.5	95.5

Table 2.26

(Given) Frequency | Bucket Midpt

Number of Stude		Bucket Midpt		
1	*	53.25	=	53.25
0	*	59.5	=	0
4	*	65.5	=	262
4	*	71.5	=	286
2	*	77.5	=	155
3	*	83.5	=	250.5
4	*	89.5	=	358
1	*	95.5	=	95.5
+ $n = 19$			+	1460.25

- Calculate the sum of the product of each interval frequency and midpoint. $\sum fm$

$$53.25(1) + 59.5(0) + 65.5(4) + 71.5(4) + 77.5(2) + 83.5(3) + 89.5(4) + 95.5(1) = 1460.25$$

- $\mu = \frac{\sum fm}{\sum f} = \frac{1460.25}{19} = 76.86$

"mu"

(population) mean $\mu \approx \frac{1460.25}{19} = \boxed{76.86}$

because the data was for the whole population